(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 896 322 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication
10.02.1999 Bulletin 1999/06

(51) Int Cl 6. G10L 9/20

(21) Application number 98306215.9

(22) Date of filing 04.08.1998

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE
Designated Extension States
AL LT LV MK RO SI

(30) Priority 05.08.1997 US 905931

(71) Applicant AT&T Corp.
New York, NY 10013-2412 (US)

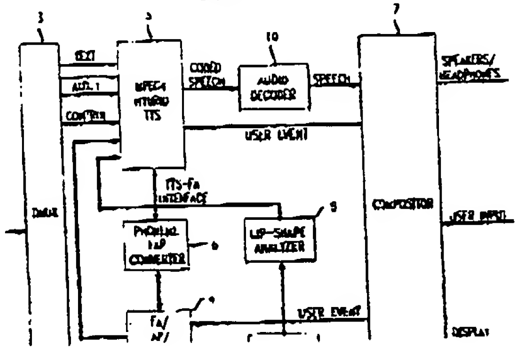
(72) Inventors.
- Basso, Andrea
N. Long Branch, New Jersey 07740 (US)
- Beutnagel, Mark Charles
Mendham, New Jersey 07945 (US)
- Ostermann, Joern
Red Bank, New Jersey 07701 (US)

(74) Representative. Asquith, Julian Peter et al
Marks & Clerk,
4220 Nash Court,
Oxford Business Park South
Oxford OX4 2RU (GB)

(54) Method and system for aligning natural and synthetic video to speech synthesis

(57) According to MPEG-4's TTS architecture, facial animation can be driven by two streams simultaneously - text, and Facial Animation Parameters In this architecture, text input is sent to a Text-To-Speech converter at a decoder that drives the mouth shapes of the face. Facial Animation Parameters are sent from an encoder to the face over the communication channel The present invention includes codes (known as bookmarks) in the text string transmitted to the Text-to-Speech converter, which bookmarks are placed between words as well as inside them According to the present invention, the bookmarks carry an encoder time stamp Due to the nature of text-to-speech conversion, the encoder time stamp does not relate to real-world time, and should be interpreted as a counter In addition, the Facial Animation Parameter stream carries the same encoder time stamp found in the bookmark of the text The system of the present invention reads the bookmark and provides the encoder time stamp as well as a real-time time stamp to the facial animation system. Finally, the facial animation system associates the correct facial animation parameter with the real-time time stamp using the encoder time stamp of the bookmark as a reference

FIG. 2



Description

BACKGROUND OF THE INVENTION

[0001] The present invention relates generally to methods and systems for coding of images, and more particularly to a method and system for coding images of facial animation.

[0002] According to MPEG-4's TTS architecture, facial animation can be driven by two streams simultaneously - text, and Facial Animation Parameters (FAPs). In this architecture, text input is sent to a Text-To-Speech (TTS) converter at a decoder that drives the mouth shapes of the face. FAPs are sent from an encoder to the face over the communication channel. Currently, the Verification Model (VM) assumes that synchronization between the input side and the FAP input stream is obtained by means of timing injected at the transmitter side. However, the transmitter does not know the timing of the decoder TTS. Hence, the encoder cannot specify the alignment between synthesized words and the facial animation. Furthermore, timing varies between different TTS systems. Thus, there currently is no method of aligning facial mimics (e.g., smiles, and expressions) with speech.

[0003] The present invention is therefore directed to the problem of developing a system and method for coding images for facial animation that enables alignment of facial mimics with speech generated at the decoder.

SUMMARY OF THE INVENTION

[0004] The present invention solves this problem by including codes (known as bookmarks) in the text string transmitted to the Text-to-Speech (TTS) converter, which bookmarks can be placed between words as well as inside them. According to the present invention, the bookmarks carry an encoder time stamp (ETS). Due to the nature of text-to-speech conversion, the encoder time stamp does not relate to real-world time, and should be interpreted as a counter. In addition, according to the present invention, the Facial Animation Parameter (FAP) stream carries the same encoder time stamp found in the bookmark of the text. The system of the present invention reads the bookmark and provides the encoder time stamp as well as a real-time time stamp (RTS) derived from the timing of its TTS converter to the facial animation system. Finally, the facial animation system associates the correct facial animation parameter with the real-time time stamp using the encoder time stamp of the bookmark as a reference. In order to prevent conflicts between the encoder time stamps and the real-time time stamps, the encoder time stamps have to be chosen such that a wide range of decoders can operate.

[0005] Therefore, in accordance with the present invention, a method for encoding a facial animation including a text stream, comprises the steps of assigning a predetermined code to the at least one facial mimic, and placing the predetermined code within the text stream, wherein said code indicates a presence of a particular facial mimic. The predetermined code is a unique escape sequence that does not interfere with the normal operation of a text-to-speech synthesizer.

[0006] One possible embodiment of this method uses the predetermined code as a pointer to a stream of facial mimics thereby indicating a synchronization relationship between the text stream and the facial mimic stream.

[0007] One possible implementation of the predetermined code is an escape sequence, followed by a plurality of bits, which define one of a set of facial mimics. In this case, the predetermined code can be placed in between words in the text stream, or in between letters in the text stream.

[0008] Another method according to the present invention for encoding a facial animation includes the steps of creating a text stream, creating a facial mimic stream, inserting a plurality of pointers in the text stream pointing to a corresponding plurality of facial mimics in the facial mimic stream, wherein said plurality of pointers establish a synchronization relationship with said text and said facial mimics.

[0009] According to the present invention, a method for decoding a facial animation including speech and at least one facial mimic includes the steps of monitoring a text stream for a set of predetermined codes corresponding to a set of facial mimics, and sending a signal to a visual decoder to start a particular facial mimic upon detecting the presence of one of the set of predetermined codes.

[0010] According to the present invention, an apparatus for decoding an encoded animation includes a demultiplexer receiving the encoded animation, outputting a text stream and a facial animation parameter stream, wherein said text stream includes a plurality of codes indicating a synchronization relationship with a plurality of mimics in the facial animation parameter stream and the text in the text stream, a text to speech converter coupled to the demultiplexer, converting the text stream to speech, outputting a plurality of phonemes, and a plurality of real-time time stamps and the plurality of codes in a one-to-one correspondence, whereby the plurality of real-time time stamps and the plurality of codes indicate a synchronization relationship between the plurality of mimics and the plurality of phonemes, and a phoneme to video converter being coupled to the text to speech converter, synchronizing a plurality of facial mimics with the plurality of phonemes based on the plurality of real-time time stamps and the plurality of codes.

[0011] In the above apparatus, it is particularly advantageous if the phoneme to video converter includes a facial animator creating a wireframe image based on the synchronized plurality of phonemes and the plurality of facial mimics, and a visual decoder being coupled to the video image based on the wireframe image

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] FIG 1 depicts the environment in which the present invention will be applied

[0013] FIG 2 depicts the architecture of an MPEG-4 decoder using text-to-speech conversion.

DETAILED DESCRIPTION

[0014] According to the present invention, the synchronization of the decoder system can be achieved by using local synchronization by means of event buffers at the input of FA/AP/MP and the audio decoder. Alternatively, a global synchronization control can be implemented

[0015] A maximum drift of 80 msec between the encoder time stamp (ETS) in the text and the ETS in the Facial Animation Parameter (FAP) stream is tolerable.

[0016] One embodiment for the syntax of the bookmarks when placed in the text stream consists of an escape signal followed by the bookmark content, e g , \!M {bookmark content}. The bookmark content carries a 16-bit integer time stamp ETS and additional information The same ETS is added to the corresponding FAP stream to enable synchronization The class of Facial Animation Parameters is extended to carry the optional ETS

[0017] If an absolute clock reference (PCR) is provided, a drift compensation scheme can be implemented Please note, there is no master slave notion between the FAP stream and the text This is because the decoder might decide to vary the speed of the text as well as a variation of facial animation might become necessary, if an avatar reacts to visual events happening in its environment.

[0018] For example, if Avatar 1 is talking to the user. A new Avatar enters the room A natural reaction of avatar 1 is to look at avatar 2, smile and while doing so, slowing down the speed of the spoken text

Autonomous Animation Driven Mostly by Text

[0019] In the case of facial animation driven by text, the additional animation of the face is mostly restricted to events that do not have to be animated at a rate of 30 frames per second Especially high-level action units like smile should be defined at a much lower rate. Furthermore, the decoder can do the interpolation between different action units without tight control from the receiver

[0020] The present invention includes action units to be animated and their intensity in the additional information of the bookmarks The decoder is required to interpolate between the action units and their intensities between consecutive bookmarks

imations using simple tools, such as text editors, and significant savings in bandwidth

[0022] FIG 1 depicts the environment in which the present invention is to be used The animation is created and coded in the encoder section 1 The encoded animation is then sent through a communication channel (or storage) to a remote destination At the remote destination, the animation is recreated by the decoder 2. At this stage, the decoder 2 must synchronize the facial animations with the speech of the avatar using only information encoded with the original animation

[0023] FIG 2 depicts the MPEG-4 architecture of the decoder, which has been modified to operate according to the present invention. The signal from the encoder 1 (not shown) enters the Demultiplexer (DMUX) 3 via the transmission channel (or storage, which can also be modeled as a channel) The DMUX 3 separates outs the text and the video data, as well as the control and auxiliary information The FAP stream, which includes the Encoder Time Stamp (ETS), is also output by the DMUX 3 directly to the FA/AP/MP 4, which is coupled to the Text-to-Speech Converter (TTS) 5, a Phoneme FAP converter 6, a compositor 7 and a visual decoder 8 A Lip Shape Analyzer 9 is coupled to the visual decoder 8 and the TTS 5. User input enters via the compositor 7 and is output to the TTS 5 and the FA/AP/MP 4. These events include start, stop, etc

[0024] The TTS 4 reads the bookmarks, and outputs the phonemes along with the ETS as well as with a Real-time Time Stamp (RTS) to the Phoneme FAP Converter 6. The phonemes are used to put the vertices of the wireframe in the correct places At this point the image is not rendered

[0025] This data is then output to the visual decoder 8, which renders the image, and outputs the image in video form to the compositor 7 It is in this stage that the FAPs are aligned with the phonemes by synchronizing the phonemes with the same ETS/RTS combination with the corresponding FAP with the matching ETS.

[0026] The text input to the MPEG-4 hybrid text-to-speech (TTS) converter 5 is output as coded speech to an audio decoder 10. In this system, the audio decoder 10 outputs speech to the compositor 7, which acts as the interface to the video display (not shown) and the speakers (not shown), as well as to the user

[0027] On the video side, video data output by the DMUX 3 is passed to the visual decoder 8, which creates the composite video signal based on the video data and the output from the FA/AP/MP 4

[0028] There are two different embodiments of the present invention. In a first embodiment, the ETS placed in the text stream includes the facial animation. That is, the bookmark (escape sequence) is followed by a 16 bit codeword that represents the appropriate facial animation to be synchronized with the speech at this point in the animation

[0029] Alternatively, the ETS placed in the text stream tion in the FAP stream. Specifically, the escape sequence is followed by a 16 bit code that uniquely identifies a particular place in the FAP stream.

[0030] While the present invention has been described in terms of animation data, the animation data could be replaced with natural audio or video data. More specifically, the above description provides a method and system for aligning animation data with text-to-speech data. However, the same method and system applies if the text-to-speech data is replaced with audio or video. In fact, the alignment of the two data streams is independent of the underlying data, at least with regard to the TTS stream.

Claims

1. A method for encoding a facial animation including at least one facial mimic and speech in the form of a text stream, comprising the steps of

   a) assigning a predetermined code to the at least one facial mimic, and
   b) placing the predetermined code within the text stream, wherein said code indicates a presence of a particular facial mimic

2. The method according to claim 1, wherein the predetermined code acts as a pointer to a stream of facial mimics thereby indicating a synchronization relationship between the text stream and the facial mimic stream

3. The method according to claim 1, wherein the predetermined code comprises an escape sequence followed by a plurality of bits, which define one of a set of possible facial mimics

4. The method according to claim 1, further comprising the step of placing the predetermined code in between words in the text stream.

5. The method according to claim 1, further comprising the step of placing the predetermined code in between letters in the text stream

6. The method according to claim 1, further comprising the step of placing the predetermined code inside words in the text stream.

7. A method for encoding a facial animation comprising the steps of

   a) creating a data stream,
   b) creating a facial mimic stream;
   c) inserting a plurality of pointers in the data stream pointing to a corresponding plurality of in said plurality of pointers establish a synchronization relationship with said data and said facial mimics

8. The method according to claim 7, wherein each of the plurality of pointers comprises a time stamp.

9. The method according to claim 7, wherein the data stream comprises a text stream that is to be converted to speech in a decoding process

10. The method according to claim 9 further comprising the step of placing at least one of the plurality of pointers in between words in the text stream

11. The method according to claim 9 further comprising the step of placing at least one of the plurality of pointers in between syllables in the text stream.

12. The method according to claim 7 further comprising the step of placing at least one of the plurality of pointers inside words in the text stream

13. The method according to claim 7, wherein the data stream comprises a video stream

14. The method according to claim 7, wherein the data stream comprises an audio stream

15. A method for decoding a facial animation including speech and at least one facial mimic, comprising the steps of:

   a) monitoring a text stream for a set of predetermined codes corresponding to a set of facial mimics; and
   b) sending a signal to a visual decoder to start a particular facial mimic upon detecting the presence of one of the set of predetermined codes

16. The method according to claim 15, wherein the predetermined code acts as a pointer to a stream of facial mimics thereby indicating a synchronization relationship between the text stream and the facial mimic stream

17. The method according to claim 15, wherein the predetermined code comprises an escape sequence

18. The method according to claim 15, further comprising the step of placing the predetermined code in between words in the text stream

19. The method according to claim 15, further comprising the step of placing the predetermined code in between phonemes in the text stream

20. The method according to claim 15, further comprising the step of placing the predetermined code inside words in the text stream.

21. An apparatus for decoding an encoded animation comprising

   a) a demultiplexer receiving the encoded animation, outputting a text stream and a facial animation parameter stream, wherein said text stream includes a plurality of codes indicating a synchronization relationship with a plurality of mimics in the facial animation parameter stream and the text in the text stream,
   b) a text to speech converter coupled to the demultiplexer, converting the text stream to speech, outputting a plurality of phonemes, and a plurality of real-time time stamps and the plurality of codes in a one-to-one correspondence, whereby the plurality of real-time time stamps and the plurality of codes indicate a synchronization relationship between the plurality of mimics and the plurality of phonemes, and
   c) a phoneme to video converter being coupled to the text to speech converter, synchronizing a plurality of facial mimics with the plurality of phonemes based on the plurality of real-time time stamps and the plurality of codes

22. The apparatus according to claim 21, further comprising a compositor converting the speech and video to a composite video signal

23. The apparatus according to claim 21, wherein the phoneme to video converter includes

   a) a facial animator creating a wireframe image based on the synchronized plurality of phonemes and the plurality of facial mimics, and
   b) a visual decoder being coupled to the demultiplexer and the facial animator, and rendering the video image based on the wireframe image

FIG. 1

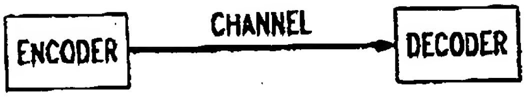
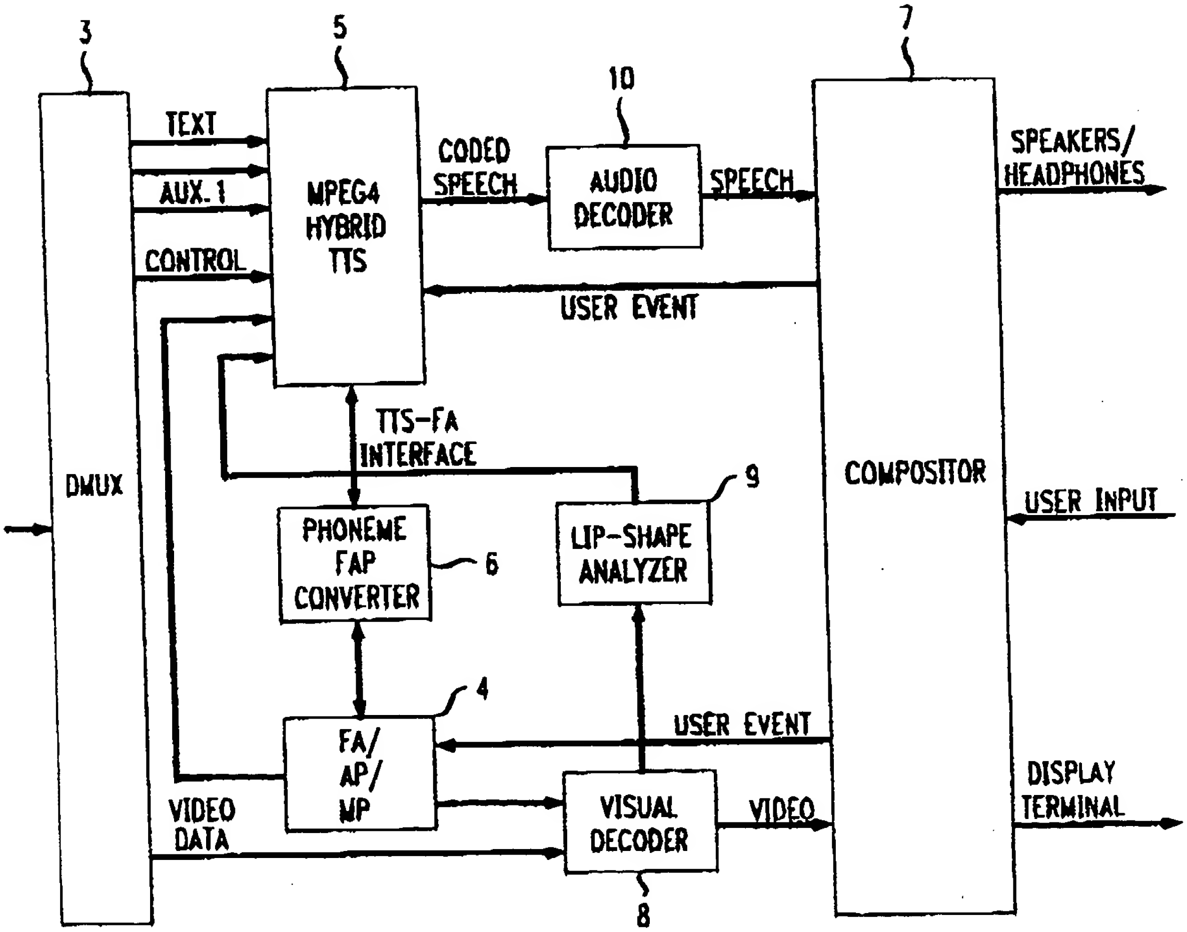

ENCODER
CHANNEL
DECODER

FIG. 2


3
5
10
7
DMUX
TEXT
AUX. 1
CONTROL
MPEG4 HYBRID TTS
CODED SPEECH
AUDIO DECODER
SPEECH
SPEAKERS/ HEADPHONES
USER EVENT
TTS-FA INTERFACE
COMPOSITOR
9
USER INPUT
PHONEME FAP CONVERTER
6
LIP-SHAPE ANALYZER
4
USER EVENT
FA/ AP/ MP
VIDEO DATA
VISUAL DECODER
VIDEO
DISPLAY TERMINAL
8

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 896 322 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
06.10.1999 Bulletin 1999/40

(43) Date of publication A2
10.02.1999 Bulletin 1999/06

(21) Application number. 98306215.9

(22) Date of filing. 04.08.1998

(51) Int Cl 6 G10L 5/02, G06T 15/70

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE
Designated Extension States
AL LT LV MK RO SI

(30) Priority 05.08.1997 US 905931

(71) Applicant AT&T Corp.
New York, NY 10013-2412 (US)

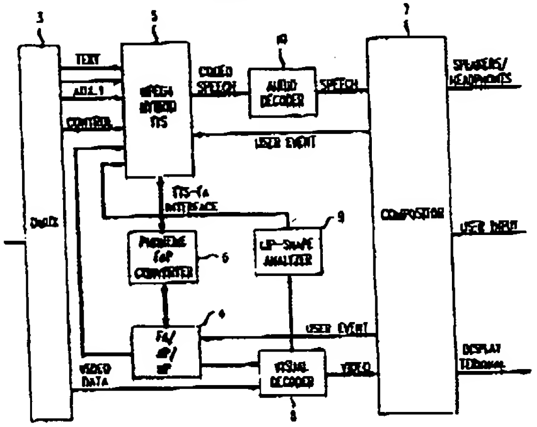
(72) Inventors.
- Basso, Andrea
  N. Long Branch, New Jersey 07740 (US)
- Beutnagel, Mark Charles
  Mendham, New Jersey 07945 (US)
- Ostermann, Joern
  Red Bank, New Jersey 07701 (US)

(74) Representative: Asquith, Julian Peter et al
Marks & Clerk,
4220 Nash Court,
Oxford Business Park South
Oxford OX4 2RU (GB)

(54) Method and system for aligning natural and synthetic video to speech synthesis

(57) According to MPEG-4's TTS architecture, facial animation can be driven by two streams simultaneously - text, and Facial Animation Parameters. In this architecture, text input is sent to a Text-To-Speech converter at a decoder that drives the mouth shapes of the face Facial Animation Parameters are sent from an encoder to the face over the communication channel. The present invention includes codes (known as bookmarks) in the text string transmitted to the Text-to-Speech converter. which bookmarks are placed between words as well as inside them According to the present invention, the bookmarks carry an encoder time stamp. Due to the nature of text-to-speech conversion, the encoder time stamp does not relate to real-world time, and should be interpreted as a counter In addition, the Facial Animation Parameter stream carries the same encoder time stamp found in the bookmark of the text The system of the present invention reads the bookmark and provides the encoder time stamp as well as a real-time time stamp to the facial animation system Finally, the facial animation system associates the correct facial animation parameter with the real-time time stamp using the encoder time stamp of the bookmark as a reference

FIG. 2



European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 98 30 6215

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X	US 4 841 575 A (WELSH WILLIAM J ET AL) 20 June 1989 (1989-06-20) * column 9, line 59 - column 10, line 27 *	1,2,7,9,15,16	G10L5/02 G06T15/70
A		21	
P,X	UZ B ET AL: "Realistic speech animation of synthetic faces" PROCEEDINGS COMPUTER ANIMATION '98, PHILADELPHIA, PA, USA, 8 - 10 June 1998, pages 111-118, XP002111637 IEEE Comput. Soc., Los Alamitos, CA, USA ISBN: 0-8186-8541-7 * section 6 ('Synchronizing Speech with Expressions), pages 115-116 *	1,2,7,9,15,16	
P,A		21	
P,X	ISO/IEC JTC 1/SC 29/WG 11: "Report of the 43rd WG 11 meeting" CODING OF MOVING PICTURES AND AUDIO. ISO/IEC JTC 1/SC 29/WG 11 N2114, March 1998 (1998-03), XP002111638 INTERNATIONAL ORGANISATION FOR STANDARDISATION * page 40, TTSI section *	1,2,7,9,15,16	**TECHNICAL FIELDS SEARCHED (Int.Cl.6)**
P,A		21	G10L G06T H04N
A	CHIARIGLIONE L: "MPEG AND MULTIMEDIA COMMUNICATIONS" IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, vol. 7, no. 1, 1 February 1997 (1997-02-01), pages 5-18, XP000678876 ISSN: 1051-8215 * sections VII ('MPEG-4 OR MULTIMEDIA COMMUNICATIONS') and VIII ('THE MPEG-4 STANDARD'), pages 12-16 *	1-23	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	9 August 1999	Ramos Sánchez, U

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EP 0 896 322 A3

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 98 30 6215

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.
09-08-1999

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 4841575 A	20-06-1989	AT 72083 T	15-02-1992
		CA 1263187 A	21-11-1989
		DE 3683609 A	05-03-1992
		EP 0225729 A	16-06-1987
		GR 3004011 T	31-03-1993
		HK 128696 A	26-07-1996
		JP 2589478 B	12-03-1997
		JP 62120179 A	01-06-1987
		JP 2753599 B	20-05-1998
		JP 8237655 A	13-09-1996

For more details about this annex : see Official Journal of the European Patent Office, No 12/82